AFRL-RI-RS-TR-2015-180

MOBILE ACTIVE AUTHENTICATION VIA LINGUISTIC MODALITIES

DREXEL UNIVERSITY

JULY 2015

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the DARPA Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2015-180 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/ S /

DANIELLE M. GAMBINO
Work Unit Manager

/ S /

WARREN H. DEBANY, JR.
Technical Advisor, Information
Exploitation and Operations Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
JULY 2015	FINAL TECHNICAL REPORT	SEP 2013 – DEC 2014

4. TITLE AND SUBTITLE

MOBILE ACTIVE AUTHENTICATION VIA LINGUISTIC MODALITIES

5a. CONTRACT NUMBER
FA8750-13-C-0268

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
62722F

6. AUTHOR(S)

Rachel Greenstadt, Moshe Kam, Lex Fridman, Patrick Brenna, John

5d. PROJECT NUMBER
AAP2

5e. TASK NUMBER
DR

5f. WORK UNIT NUMBER
EX

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Drexel University
3141 Chestnut St
Philadelphia, PA 19104-2875

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIGA
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2015-180

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. PA# DARPA DISTAR #24880
Date Cleared: 14 JUL 2015

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Active authentication is the problem of continuously verifying the identity of a person based on behavioral aspects of their interaction with a computing device. In this study, we collect and analyze behavioral biometrics data from 200 subjects, each using their personal Android mobile device for a period of at least 30 days. This dataset is novel in the context of active authentication due to its size, duration, number of modalities, and absence of restrictions on tracked activity. The geographical colocation of the subjects in the study is representative of a large closed-world environment such as an organization where the unauthorized user of a device is likely to be an insider threat: coming from within the organization. We consider four biometric modalities: (1) text entered via soft keyboard, (2) applications used, (3) websites visited, and (4) physical location of the device as determined from GPS (when outdoors) or WiFi (when indoors). We implement and test a classifier for each modality and organize the classifiers as a parallel binary decision fusion architecture. We are able to characterize the performance of the system with respect to intruder detection time and to quantify the contribution of each modality to the overall performance. We further characterize the contribution of two additional modalities developed in addition to the main four modalities: eye tracking and an alternate stylometry method.

15. SUBJECT TERMS

Active authentication, behavioral biometrics, mobile devices, stylometry, web browsing behavior, application usage patterns, eye tracking

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**DANIELLE M. GAMBINO**
U	U	U	UU	29	19b. TELEPHONE NUMBER (*Include area code*) **N/A**

Contents

List of Figures

List of Tables

1 Summary

According to a 2013 Pew Internet Project study of 2076 people [1], 91% of American adults own a cellphone. Increasingly, people are using their phones to access and store sensitive data. The same study found that 81% of cellphone owners use their mobile device for texting, 52% use it for email, 49% use it for maps (enabling location services), and 29% use it for online banking. And yet, securing the data is often not taken seriously because of an inaccurate estimation of risk as discussed in [2]. In particular, several studies have shown that a large percentage of smartphone owners do not lock their phone: 57% in [3], 33% in [4], 39% in [2], and 48% in this study.

Active authentication is an approach of monitoring the behavioral biometric characteristics of a user's interaction with the device for the purpose of securing the phone when the point-of-entry locking mechanism fails or is absent. In recent years, continuous authentication has been explored extensively on desktop computers, based either on a single biometric modality like mouse movement [5] or a fusion of multiple modalities like keyboard dynamics, mouse movement, web browsing, and stylometry [6]. Unlike physical biometric devices like fingerprint scanners or iris scanners, these systems rely on computer interface hardware like the keyboard and mouse that are already commonly available with most computers.

In this report, we consider the problem of active authentication on mobile devices, where the variety of available sensor data is much greater than on the desktop, but so is the variety of behavioral profiles, device form factors, and environments in which the device is used. We study four representative modalities of stylometry (text analysis), application usage patterns, web browsing behavior, and physical location of the device. In the remainder of the report these four modalities will be referred to as text, app, web, and location, respectively. We consider the trade-off between intruder detection time and detection error as measured by false accept rate (FAR) and false reject rate (FRR). The analysis is performed on a dataset collected by the authors of 200 subjects using their personal Android mobile device for a period of at least 30 days. To the best of our knowledge, this dataset is the first of its kind studied in active authentication literature, due to its large size [7], the duration of tracked activity [8], and the absence of restrictions on usage patterns and on the form factor of the mobile device. The geographical colocation of the participants, in particular, makes the dataset a good representation of an environment such as a closed-world organization where the unauthorized user of a particular device will most likely come from inside the organization.

We propose to use decision fusion in order to asynchronously integrate the four modalities and make serial authentication decisions. While we consider here a specific set of binary classifiers, the strength of our decision-level approach is that additional classifiers can be added without having to change the basic fusion rule. Moreover, it is easy to evaluate the marginal improvement of any added classifier to the overall performance of the system. We evaluate the multimodal continuous authentication system by characterizing the error rates of local classifier decisions, fused global decisions, and the contribution of each local classifier to the fused decision. The novel aspects of our work include the scope of the dataset, the particular portfolio of behavioral biometrics in the context of mobile devices, and the extent of temporal performance analysis.

The remainder of the report is structured as follows. In §2, we discuss the related work on multimodal biometric systems, active authentication on mobile devices, and each of the four behavioral biometrics considered in this report. In §3, we discuss the 200 subject dataset that we collected and analyzed. In §4, we discuss four biometric modalities, their associated classifiers, and the decision fusion architecture. In §5, we present the performance of each individual classifier, the performance of the fusion system, and the contribution of each individual classifier to the fused decisions.

2 Introduction

2.1 Multimodal Biometric Systems

The window of time based on which an active authentication system is tasked with making a binary decision is relatively short and thus contains a highly variable set of biometric information. Depending on the task the user is engaged in, some of the biometric classifiers may provide more data than others. For example, as the user chats with a friend via SMS, the text-based classifiers will be actively flooded with data, while the web browsing based classifiers may only get a few infrequent events. This motivates the recent work on multimodal authentication systems where the decisions of multiple classifiers are fused together [9]. In this way, the verification process is more robust to the dynamic nature of human-computer interaction. The current approaches to the fusion of classifiers center around max, min, median, or majority vote combinations [10]. When neural networks are used as classifiers, an ensemble of classifiers is constructed and fused based on different initialization of the neural network [11].

Several active authentication studies have utilized multimodal biometric systems but have all, to the best of our knowledge: (1) considered a smaller pool of subjects, (2) have not characterized the temporal performance of intruder detection, and (3) have shown overall significantly worse performance than that achieved in our study.

Our approach in this report is to apply the Chair-Varshney optimal fusion rule [12] for the combination of available multimodal decisions. The strength of the decision-level fusion approach is that an arbitrary number of classifiers can be added without re-training the classifiers already in the system. This modular design allows for multiple groups to contribute drastically different classification schemes, each lowering the error rate of the global decision.

2.2 Mobile Active Authentication

With the rise of smartphone usage, active authentication on mobile devices has begun to be studied in the last few years. The large number of available sensors makes for a rich feature space to explore. Ultimately, the question is the one that we ask in this report: what modality contributes the most to a decision fusion system toward the goal of fast, accurate verification of identity? Most of the studies focus on a single modality. For example, gait pattern was considered in [7] achieving an EER of 0.201 (20.1%) for 51 subjects during two short sessions, where each subject was tasked with walking down a hallway. Some studies have incorporated multiple modalities. For example, keystroke dynamics, stylometry, and behavioral profiling were

considered in [13] achieving an EER of 0.033 (3.3%) from 30 simulated users. The data for these users was pieced together from different datasets. To the best of our knowledge, the dataset that we collected and analyzed is unique in all its key aspects: its size (200 subjects), its duration (30+ days), and the size of the portfolio of modalities that were all tracked concurrently with a synchronized timestamp.

2.3 Stylometry, Web Browsing, Application Usage, Location

Stylometry is the study of linguistic style. It has been extensively applied to the problems of authorship attribution, identification, and verification. See [14] for a thorough summary of stylometric studies in each of these three problem domains along with their study parameters and the resulting accuracy. These studies traditionally use large sets of features (see Table II in [15]) in combination with support vector machines (SVMs) that have proven to be effective in high dimensional feature space [16], even in cases when the number of features exceeds the number of samples. Nevertheless, with these approaches, often more than 500 words are required in order to achieve adequately low error rates [17]. This makes them impractical for the application of real-time active authentication on mobile devices where text data comes in short bursts.

While the other three modalities are not well investigated in the context of active authentication, this is not true for stylometry. Therefore, for this modality, we don't reinvent the wheel, and implement the n-gram analysis approach presented in [14] that has been shown to work sufficiently well on short blocks of texts.

Web browsing, application usage, and location have not been studied extensively in the context of active authentication. The following is a discussion of the few studies that we are aware of. Web browsing behavior has been studied for the purpose of understanding user behavior, habits, and interests [18]. Web browsing as a source for behavioral biometric data was considered in [19] to achieve average identification FAR/FRR of 0.24 (24%) on a dataset of 14 desktop computer users. Application usage was considered in [8], where cellphone data (from 2004) from the MIT Reality Mining project [20] was used to achieve 0.1 (10%) EER based on a portfolio of metrics including application usage, call patterns, and location. Application usage and movements patterns have been studied as part of behavioral profiling in cellular networks [8,21,22]. However, these approaches use position data of lower resolution in time and space than that provided by GPS on smartphones. To the best of our knowledge, GPS traces have not been utilized in literature for continuous authentication.

3 Methods

The dataset used in this work contains behavioral biometrics data for 200 subjects. The collection of the data was carried out by the authors over a period of 5 months. The requirements of the study were that each subject was a student or employee of Drexel University and was an owner and an active user of an Android smartphone or tablet. The number of subjects with each major Android version and associated API level are listed in Table 1. Nexus 5 was the most popular

device with 10 subjects using it. Samsung Galaxy S5 was the second most popular device with 6 subjects using it.

Table 1: The Android version and API level of the 200 devices that were part of the study.

Android Version	API Level	Subjects
4.4	19	143
4.1	16	16
4.3	18	15
4.2	17	9
4.0.4	15	5
2.3.6	10	4
4.0.3	15	3
2.3.5	10	3
2.2	8	2

A tracking application was installed on each subject's device and operated for a period of at least 30 days until the subject came in to approve the collected data and get the tracking application uninstalled from their device. The following data modalities were tracked with 1-second resolution:

- Text typed via soft keyboard.
- Apps visited.
- Websites visited.
- Location (based on GPS or WiFi).

The key characteristics of this dataset are its large size (200 users), the duration of tracked activity (30+ days), and the geographical colocation of its participants in the Philadelphia area. Moreover, we did not place any restrictions on usage patterns, on the type of Android device, and on the Android OS version (see Table 1).

There were several challenges encountered in the collection of the data. The biggest problem was battery drain. Due to the long duration of the study, we could not enable modalities whose tracking proved to be significantly draining of battery power. These modalities include front-facing video for eye tracking and face recognition, gyroscope, accelerometer, and touch gestures. Moreover, we had to reduce GPS sampling frequency to once per minute on most of the devices.

Table 2: The number of events in the dataset associated with each of the four modalities considered in this report.

Event	Frequency
Text	23,254,478
App	927,433
Web	210,322
Location	143,875

A text event refers to a single character entered on the soft keyboard. An app events refers to a new app receiving focus. A web event refers to a new url entered in the url box. A location event refers to a new sample of the device location either from GPS or WiFi.

Table 2 shows statistics on each of the four investigated modalities in the corpus. The table contains data aggregated over all 200 users. The "frequency" here is a count of the number of instances of an action associated with that modality. As stated previously, the four modalities will be referred to as text, app, web, and "location." For text, the action is a single keystroke on the soft keyboard. For app, the action is opening or bringing focus to a new app. For web, the action is visiting a new website. For location, no explicitly action is taken by the user. Rather, location is sampled regularly at intervals of 1 minute when GPS is enabled. As Table 2 suggests, text events fire 1-2 orders of magnitude more frequently than the other three.

The data for each user is processed to remove idle periods when the device is not active. The threshold for what is considered an idle period is 5 minutes. For example, if the time between event A and event B is 20 minutes, with no other events in between, this 20 minutes is compressed down to 5 minutes. The date and time of the event are not changed but the timestamp used in dividing the dataset for training and testing (see §5.1) is updated to reflect the new time between event A and event B. This compression of idle times is performed in order to regularize periods of activity for cross validation that utilizes time-based windows as described in §5.1. The resulting compressed timestamps are referred to as "active interaction". Fig. 1 shows the duration (in hours) of active interaction for each of the 200 users ordered from least to most active.

Table 3 shows three top-20 lists: (1) the top-20 apps based on the amount of text that was typed inside each app, (2) the top-20 apps based on the number of times they received focused, and (3) the top-20 website domains based on the number of times a website associated with that domain was visited. These are aggregate measures across the dataset intended to provide an intuition about its structure and content, but the top-20 list is the same as that used for the the classifier model based on the web and app features in §4.

Fig. 2 shows a heat map visualization of a selection from the dataset of GPS locations in the Philadelphia area. The subjects in the study resided in Philadelphia but traveled all over United States and the world. There are two key characteristics of the GPS location data. First, it is relatively unique to each individual even for people living in the same area of a city. Second,

outside of occasional travel, it does not vary significantly from day to day. Human beings are creatures of habit, and in as much as location is a measure of habit, this idea is confirmed by the location data of the majority of the subjects in the study.

Table 3: Top 20 apps ordered by text entry and visit frequency and top 20 websites ordered by visit frequency. These tables are provided to give insight into the structure and content of the dataset.

App Name	Keys Per App
com.android.sms	5,617,297
com.android.mms	5,552,079
com.whatsapp	4,055,622
com.facebook.orca	1,252,456
com.google.android.talk	1,147,295
com.infraware.polarisviewer4	990,319
com.android.chrome	417,165
com.facebook.katana	405,267
com.snapchat.android	377,840
com.google.android.gm	271,570
com.htc.sense.mms	238,300
com.tencent.mm	221,461
com.motorola.messaging	203,649
com.android.calculator2	167,435
com.verizon.messaging.vzmsgs	137,339
com.groupme.android	134,896
com.handcent.nextsms	123,065
com.jb.gosms	118,316
com.sonyericsson.conversations	114,219
com.twitter.android	92,605

(a)

App Name	Visits
TouchWiz home	101,151
WhatsApp	64,038
Messaging	60,015
Launcher	39,113
Facebook	38,591
Google Search	32,947
Chrome	32,032
Snapchat	23,481
System UI	22,772
Phone	19,396
Gmail	19,329
Messages	19,154
Contacts	18,668
Hangouts	17,209
Home	16,775
HTC Sense	16,325
YouTube	14,552
Xperia Home	13,639
Instagram	13,146
Settings	12,675

(b)

Website Domain	Visits
www.google.com	19,004
m.facebook.com	9,300
www.reddit.com	4,348
forums.huaren.us	3,093
learn.dcollege.net	2,133
en.m.wikipedia.org	1,825
mail.drexel.edu	1,520

one.drexel.edu	1,472
login.drexel.edu	1,462
likes.com	1,361
mail.google.com	1,292
i.imgur.com	1,132
www.amazon.com	1,079
netcontrol.irt.drexel.edu	1,049
www.facebook.com	903
banner.drexel.edu	902
m.hupu.com	824
t.co	801
duapp2.drexel.edu	786
m.ign.com	725

(c)

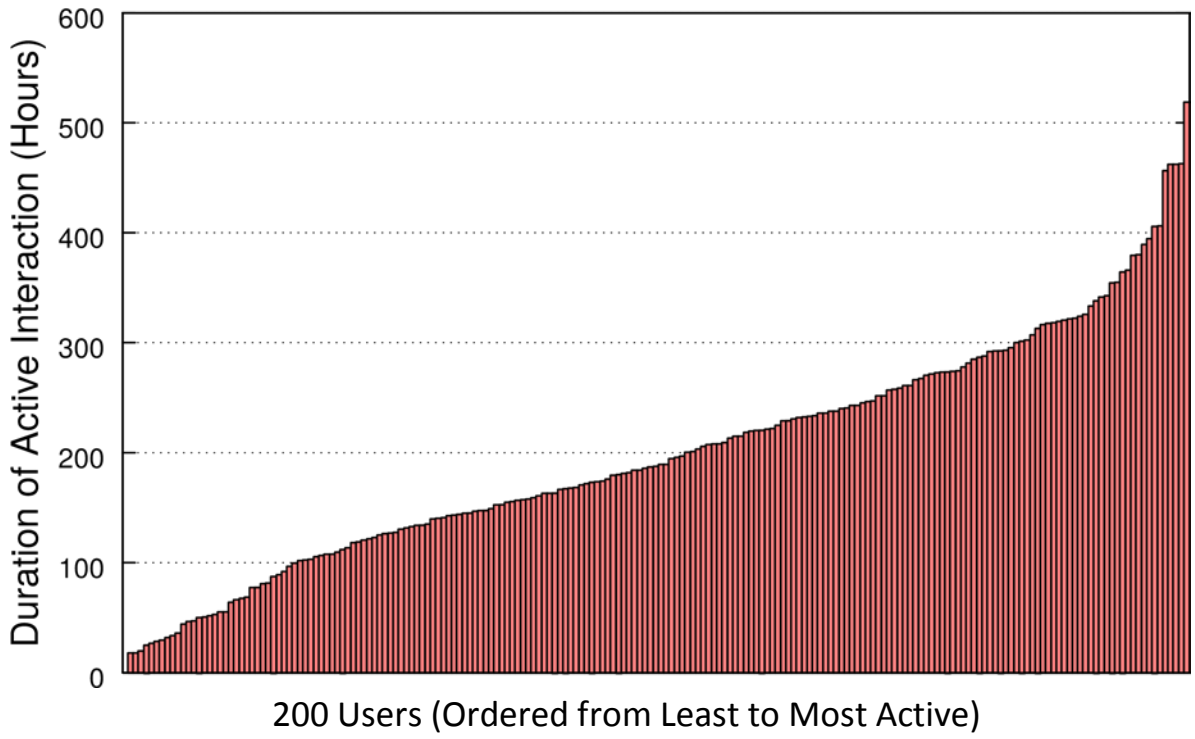

Figure 1: The duration of time (in hours) that each of the 200 users actively interacted with their device..

4 Assumptions and Procedures

4.1 Features and Classifiers

The four distinct biometric modalities considered in our analysis are (1) text entered via soft keyboard, (2) applications used, (3) websites visited, and (4) physical location of the device as determined from GPS (when outdoors) or WiFi (when indoors). We refer to these four modalities as text, app, web, and location, respectively. In this section we discuss the features that were extracted from the raw data of each modality, and the classifiers that were used to map these features into binary decision space.

A binary classifier is constructed for each of the 200 users and 4 modalities. In total, there are 800 classifiers, each producing either a probability that a user is valid $P(H_1)$ (or a binary decision of 0 (invalid) or 1 (valid). The first class (H_1) for each classifier is trained on the valid user's data and the second class (H_0) is trained on the other 199 users' data. The training process is described in more detail in §5.1. For app, web, and location, the classifier takes a single instance of the event and produces a probability. For multiple events of the same modality, the set of probabilities is fused across time using maximum likelihood:

$$H^* = \underset{i\in\{0,1\}}{\operatorname{argmax}} \prod_{x_t\in\Omega} P(x_t|H_i), \tag{1}$$

where $\Omega = \{x_t | T_{\text{current}} - T(x_t) \leq \omega\}$, ω is a fixed window size in seconds, $T(x_t)$ is the timestamp of event x_t, and T_{current} is the current timestamp. The process of fusing classifier scores across time is illustrated in Fig. 3.

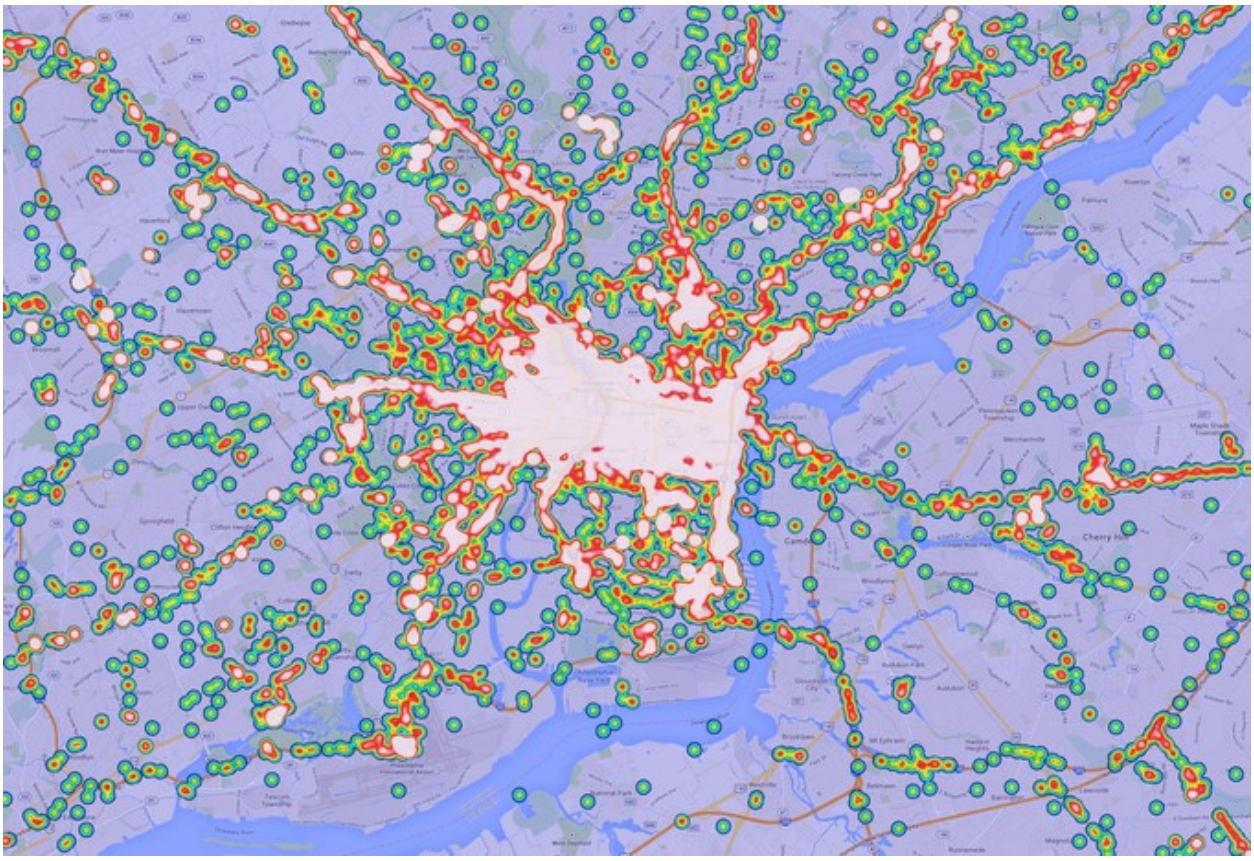

Figure 2: An aggregate heatmap showing a selection from the dataset of GPS locations in the Philadelphia area.

4.1.1 Text

As Table 3a indicates, the apps into which text was entered on mobile devices varied, but the activity in majority of the cases was communication via SMS, MMS, WhatsApp, Facebook, Google Hangouts, and other chat apps. Therefore, `text` events fired in short bursts. The tracking application captured the keys that were touched on the keyboard and not the autocorrected result. Therefore, the majority of the typed messages had a lot of misspellings and words that were erased in the final submitted message. In the case of SMS, we also were able to record the submitted result. For example, an SMS text that was submitted as "Sorry couldn't call back." had associated with it the following recorded keystrokes: "Sprry coyld cpuldn't vsll back." Classification based on the actual typed keys in principle is a better representation of the person's linguistic style. It captures unique typing idiosyncrasies that autocorrect can conceal. As discussed in §2, we implemented a one-feature n-gram classifier from [14] that has been shown to work well on short messages. It works by analyzing the presence or absence of n-grams with respect to the training set.

4.1.2 App and Web

The app and web classifier models we construct are identical in their structure. For the app modality we use the app name as the unique identifier and count the number of times a user visits each app in the training set. For the web modality we use the domain of the URL as the unique identifier and count the number of times a user visits each domain in the training set. Note that, for example, “m.facebook.com” is a considered a different domain than “www.facebook.com” because the subdomain is different. In this section we refer to the app name and the web domain as an “entity”. Table 3b and Table 3c show the top entities aggregated across all 200 users for app and web respectively.

For each user, the classification model for the valid class is constructed by determining the top 20 entities visited by that user in the training set. The quantity of visits is then normalized so that the 20 frequency values sum to 1. The classification model for the invalid class is constructed by counting the number of visit by the other 199 users to those same 20 domains, such that for each of those domains we now have a probability that a valid user visits it and an invalid user visits it. The evaluation for each user given the two empirical distributions is performed by the maximum likelihood product in (1). Entities that do not appear in the top 20 are considered outliers and are ignored in this classifier.

4.1.3 Location

Location is specified as a pair of values: latitude and longitude. Classification is performed using support vector machines (SVMs) [23] with the radial basis function (RBF) as the kernel function. The SVM produces a classification score for each pair of latitude and longitude. This score is calibrated to form a probability using Platt scaling [24] which requires an extra logistic regression on the SVM scores via an additional cross-validation on the training data. All of the code in this report is written by the authors except for the SVM classifier. Since the authentication system is written in C++, we used the Shark 3.0 machine learning library for the SVM implementation.

4.2 Decision Fusion

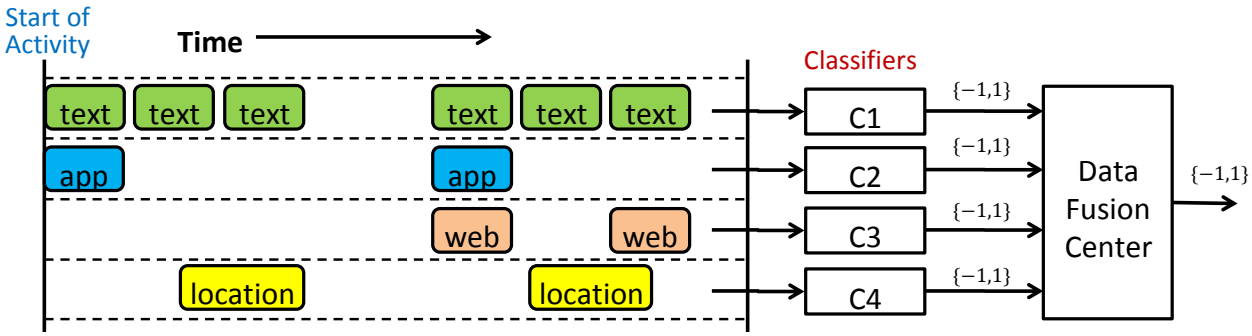

Figure 3: The fusion architecture across time and across classifiers. The text, app, web, and location boxes indicate a firing of a single event associated with each of those modalities.

Multiple classifier scores from the same modality are fused via (1) to produce a single local binary decision. Local binary decisions from each of the four modalities are fused via (4) to produce a single global binary decision.

Decision fusion with distributed sensors is described by Tenney and Sandell in [25] who studied a parallel decision architecture. As described in [26], the system comprises of n local detectors, each making a decision about a binary hypothesis (H_0,H_1), and a decision fusion center (DFC) that uses these local decisions $\{u_1,u_2,\ldots,u_n\}$ for a global decision about the hypothesis. The i^{th} detector collects K observations before it makes its decision, u_i. The decision is $u_i = 1$ if the detector decides in favor of H_1 and $u_i = -1$ if it decides in favor of H_0. The DFC collects the n decisions of the local detectors and uses them in order to decide in favor of $H_0(u = -1)$ or in favor of $H_1(u = 1)$. Tenney and Sandell [25] and Reibman and Nolte [27] studied the design of the local detectors and the DFC with respect to a Bayesian cost, assuming the observations are independent conditioned on the hypothesis. The ensuing formulation derived the local and DFC decision rules to be used by the system components for optimizing the system-wide cost. The resulting design requires the use of likelihood ratio tests by the decision makers (local detectors and DFC) in the system. However the thresholds used by these tests require the solution of a set of nonlinear coupled differential equations. In other words, the design of the local decision makers and the DFC are co-dependent. In most scenarios the resulting complexity renders the quest for an optimal design impractical.

Chair and Varshney in [12] developed the optimal fusion rule when the local detectors are fixed and local observations are statistically independent conditioned on the hypothesis. Data Fusion Center is optimal given the performance characteristics of the local fixed decision makers. The result is a suboptimal (since local detectors are fixed) but computationally efficient and scalable design. In this study we use the ChairVarshney formulation. The parallel distributed fusion scheme (see Fig. 3) allows each classifier to observe an event, minimize the local risk and make a local decision over the set of hypothesis, based on only its own observations. Each classifier sends out a decision of the form:

$$u_i = \begin{cases} 1, & \text{if } H_1 \text{ is decided} \\ -1, & \text{if } H_0 \text{ is decided} \end{cases} \tag{2}$$

The fusion center combines these local decisions by minimizing the global Bayes' risk. The optimum decision rule performs the following likelihood ratio test

$$\frac{P(u_1, ..., u_n|H_1)}{P(u_1, ..., u_n|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_0}{P_1} = \tau \tag{3}$$

where the a priori probabilities of the binary hypotheses H_1 and H_0 are P_1 and P_0 respectively. In this case the general fusion rule proposed in [12] is

$$f(u_1, ..., u_n) = \begin{cases} 1, & \text{if } a_0 + \sum_{i=0}^{n} a_i u_i > 0 \\ -1, & \text{otherwise} \end{cases} \tag{4}$$

with P_i^M,P_i^F representing the *False Rejection Rate* (FRR) and *False Acceptance Rate* (FAR) of the i^{th} classifier respectively. The optimum weights minimizing the global probability of error are given by

$$a_0 = \log \frac{P_1}{P_0}$$

$$a_i = \begin{cases} \log \frac{1-P_i^M}{P_i^F}, & \text{if } u_i = 1 \\ \log \frac{1-P_i^F}{P_i^M}, & \text{if } u_i = -1 \end{cases} \tag{5}$$

(6)

The threshold in (3) requires knowledge of the a priori probabilities of the hypotheses. In practice, these probabilities are not available, and the threshold τ is determined using different considerations such as fixing the probability of false alarm or false rejection as is done in §5.3.

5 Results and Discussion

5.1 Training, Characterization, Testing

The data of each of the 200 users' active interaction with the mobile device was divided into 5 equal-size folds (each containing 20% time span of the full set). We performed training of each classifier on the first three folds (60%). We then tested their performance on the fourth fold. This phase is referred to as "characterization", because its sole purpose is to form estimates of FAR and FRR for use by the fusion algorithm. We then tested the performance of the classifiers,

individually and as part of the fusion system, on the fifth fold. This phase is referred to as "testing" since this is the part that is used for evaluation the performance of the individual classifiers and the fusion system. The three phases of training, characterization, and testing as they relate to the data folds are shown in Fig. 4.

- Training on folds 1, 2, 3.
  Characterization on fold 4.
  Testing on fold 5.
- Training on folds 2, 3, 4.
  Characterization on fold 5.
  Testing on fold 1.
- Training on folds 3, 4, 5.
  Characterization on fold 1.
  Testing on fold 2.
- Training on folds 4, 5, 1.
  Characterization on fold 2.
  Testing on fold 3.
- Training on folds 5, 1, 2.
  Characterization on fold 3.
  Testing on fold 4.

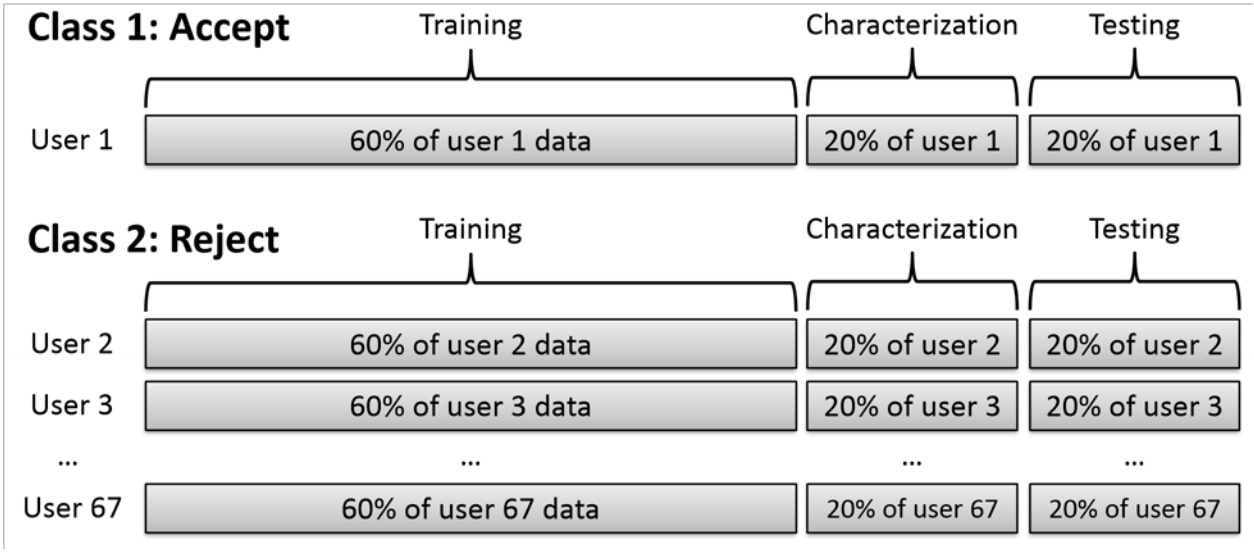

Figure 4: The three phases of processing the data to determine the individual performance of each classifiers and the performance of the fusion system that combines some subset of these classifiers.

The common evaluation method used with each classifier for data fusion was measuring the averaged error rates across five experiments; In each experiment, data of 3 folds was taken for training, 1 fold for characterization, and 1 for testing. The FAR and FRR computed during characterization were taken as input for the fusion system as a measurement of the expected performance of the classifiers. Therefore each experiment consisted of three phases: 1) train the classifier(s) using the training set, 2) determine FAR and FRR based on the training set, and 3) classify the windows in the test set.

5.2 Performance: Individual Classifiers

The conflicting objectives of an active authentication system are of response-time and performance. The less the system waits before making an authentication decision, the higher the expected rate of error. As more behavioral biometric data trickles in, the system can, on average, make a classification decision with greater certainty.

This pattern of decreased error rates with an increased decision window can be observed in Fig. 5 that shows (for 10 different time windows) the FAR and FRR of the 4 classifiers averaged over the 200 users with the error bars indicating the standard deviation. The "testing fold" (see §5.1) is used for computing these error rates. The "characterization fold" does not affect these results, but is used only for FAR/FRR estimation required by the decision fusion center in §5.3.

The "time before decision" is the time between the first event indicating activity and the first decision produced by the fusion system. This metric can be thought of as "decision window size". Events older than the time range covered by the time-window are disregarded in the classification. If no event associated with the modality under consideration fires in a specific time window, no error is added to the average.

Table 4: The rates at which an event associated with each modality "fires" per hour. On average, GPS location is provided only 3.5 times an hour.

Event	Firing Rate (per hour)
Text	557.8
App	23.2
Web	5.6
Location	3.5

There are two notable observations about the FAR/FRR plots in Fig. 5. First, the location modality provides the lowest error rates even though on average across the dataset it fires only 3.5 times an hour as shown in Table 4. This means that classification on a single GPS coordinate is sufficient to correctly verify the user with an FAR of under 0.1 and an FRR of under 0.05. Second, the text modality converges to an FAR of 0.16 and an FRR of 0.11 after 30 minutes which is one of the worse performers of the four modalities, even though it fires 557.8 times an hour on average. At the 30 minute mark, that firing rate equates to an average text block size of 279 characters. An FAR/FRR of 0.16/0.11 with 279 characters blocks improves on the error rates

achieved in [14] with 500 character blocks which in turn improved on the errors rates achieved in prior work for blocks of small text (see [14] for a full reference list on short-text stylometric analysis).

In addition to the main four features under consideration in this report, we evaluated the contribution of eye tracking and an alternate stylometry feature-set. The EER of the eye tracking metric was 0.12 and the EER of the alternate stylometry metric was 0.26.

5.3 Performance: Decision Fusion

The events associated with each of the 4 modalities fire at very different rates as shown in Table 4. Moreover, text events fire in bursts, while the location events fire at regularly spaced intervals when GPS signal is available. The app and web events fire at varying degrees of burstiness depending on the user. Fig. 6 shows the distribution of the number of events that fire within each of the time windows. An important takeaway from these distributions is that most events come in bursts followed by periods of inactivity. This results in the counterintuitive fact that the 1 minute, 10 minute, and 30 minute windows have a similar distribution on the number of events that fire within them. This is why the decrease in error rates attained from waiting longer for a decision is not as significant as might be expected.

Asynchronous fusion of classification of events from each of the four modalities is robust to the irregular rates at which events fire. The decision fusion rule in (4) utilizes all the available biometric data, weighing each classifier according to its prior performance. Fig. 7 shows the receiver operating characteristic (ROC) curve trading off between FAR and FRR by varying the threshold parameter τ in (3).

As the size of the decision window increases, the performance of the fusion system improves, dropping from an equal error rate (EER) of 0.05 using the 1 minute window to below 0.01 EER using the 30 minute window.

5.4 Contribution of Local Classifiers to Global Decision

The performance of the fusion system that utilizes all four modalities of text, app, web, and location is described in the previous section. Besides this, we are able to use the fusion system to characterize the contribution of each of the local classifiers to the global decision. This is the central question we consider in the report: what biometric modality is most helpful in verifying a person's identity under a constraint of a specific time window before the verification decision must be made? We measure the contribution C_i of each of the four classifiers by evaluating the performance of the system with and without the classifier, and computing the contribution by:

$$C_i = \frac{E_i - E}{E_i} \tag{7}$$

where E is the error rate computed by averaging FAR and FRR of the fusion system using the full portfolio of 4 classifiers, E_i is the error rate of the fusion system using all but the i-th

classifier, and C_i is the relative contribution of the i-th classifier as shown in Fig. 8. We consider the contribution of each classifier under three time windows of 1 minute, 10 minutes, and 30 minutes. Location contributes the most in all three cases, with the second biggest contributor being web browsing. Text contributes the least for the small window of 1 minute, but improve for the large windows. App usage is the least predictable contributor.

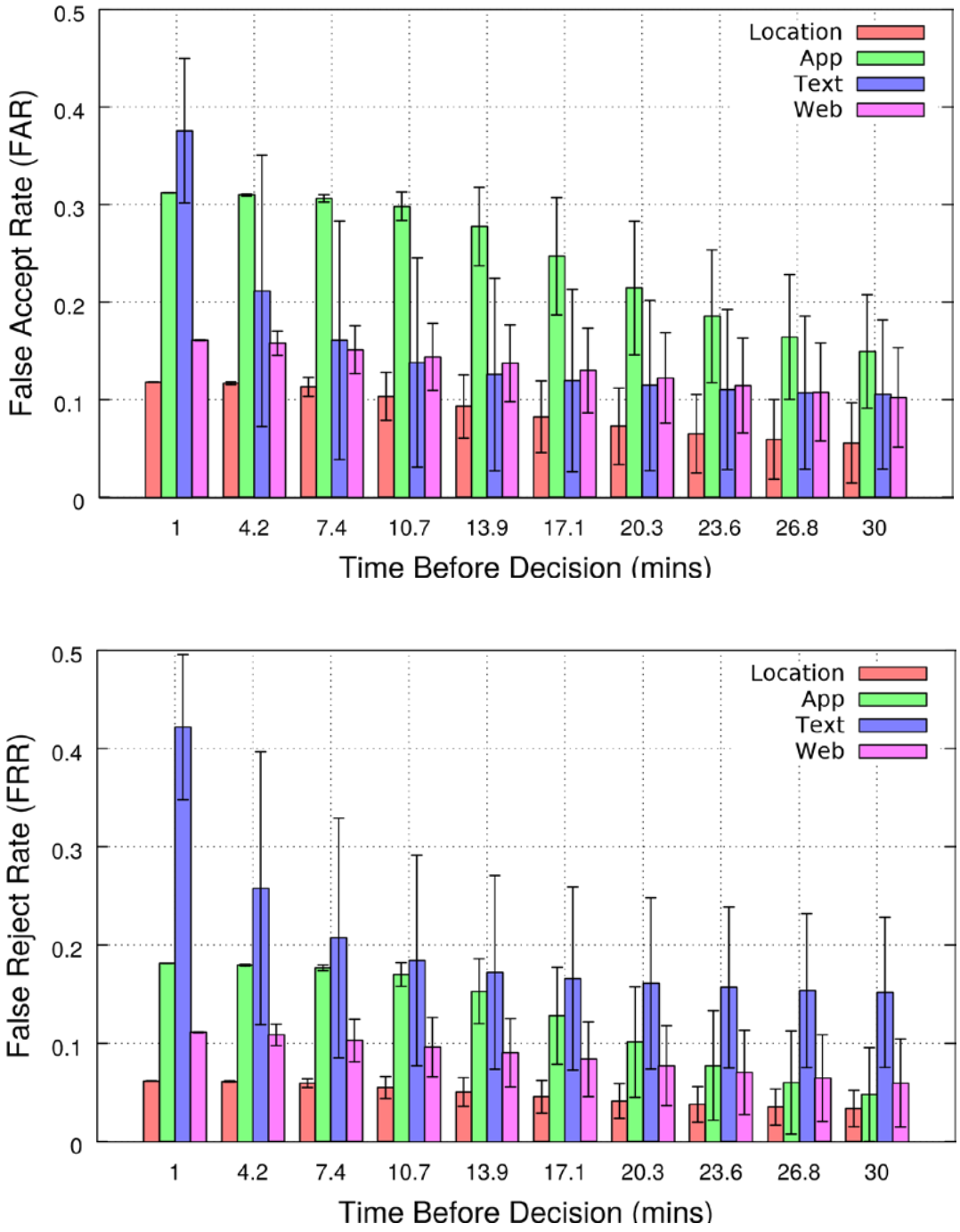

Figure 5: FAR and FRR performance of the individual classifiers associated with each of the four modalities.

Each bar represent the average error rate for a given module and time window. Each of the 200 users has 2 classifiers for each modality, so each bar provides a value that was averaged over 200 individual error rates. The error bar indicate the standard deviation across these 200 values.

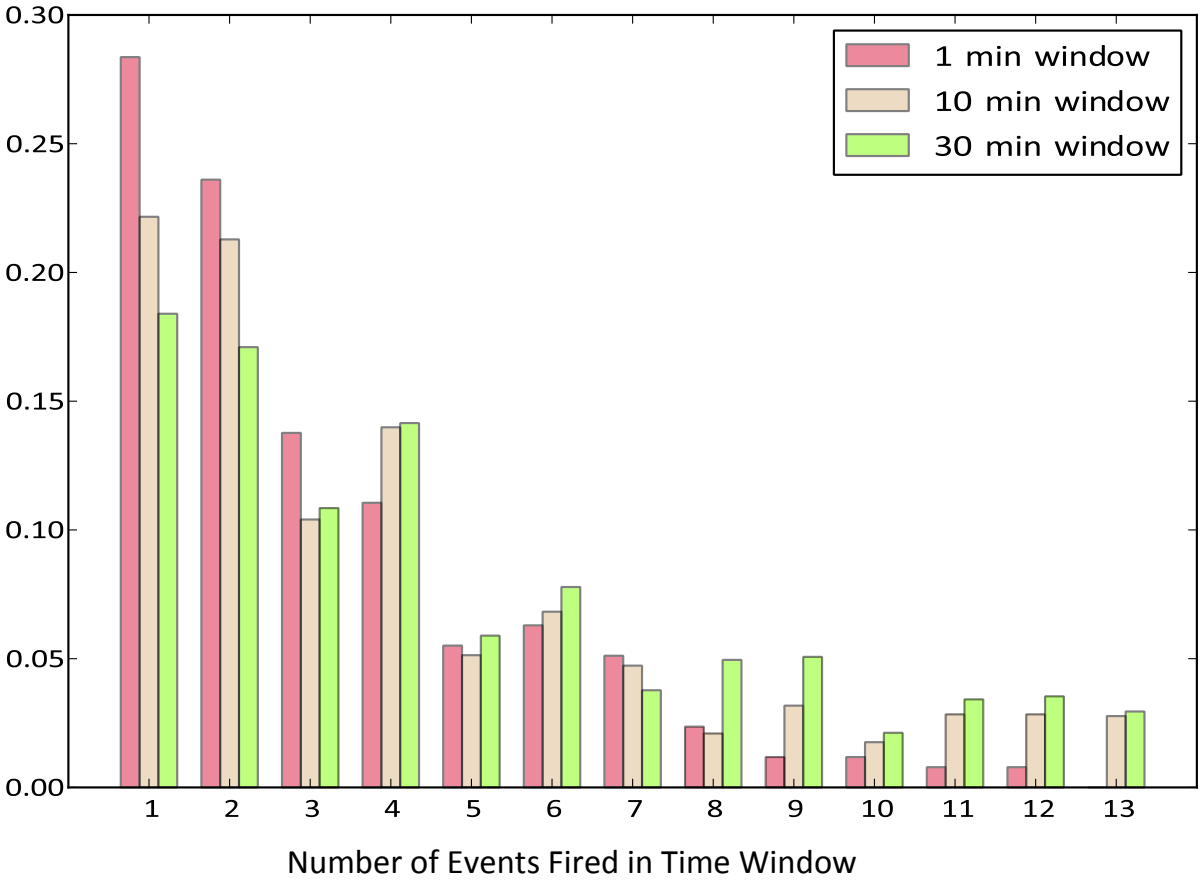

Figure 6: The distribution of the number of events that fire within a given time window.

This is a long tail distribution as non-zero probabilities of event frequencies above 13 extend to over 100. These outliers are excluded from this histogram plot in order to highlight the high-probability frequencies. Time windows in which no events fire are not included in this plot.

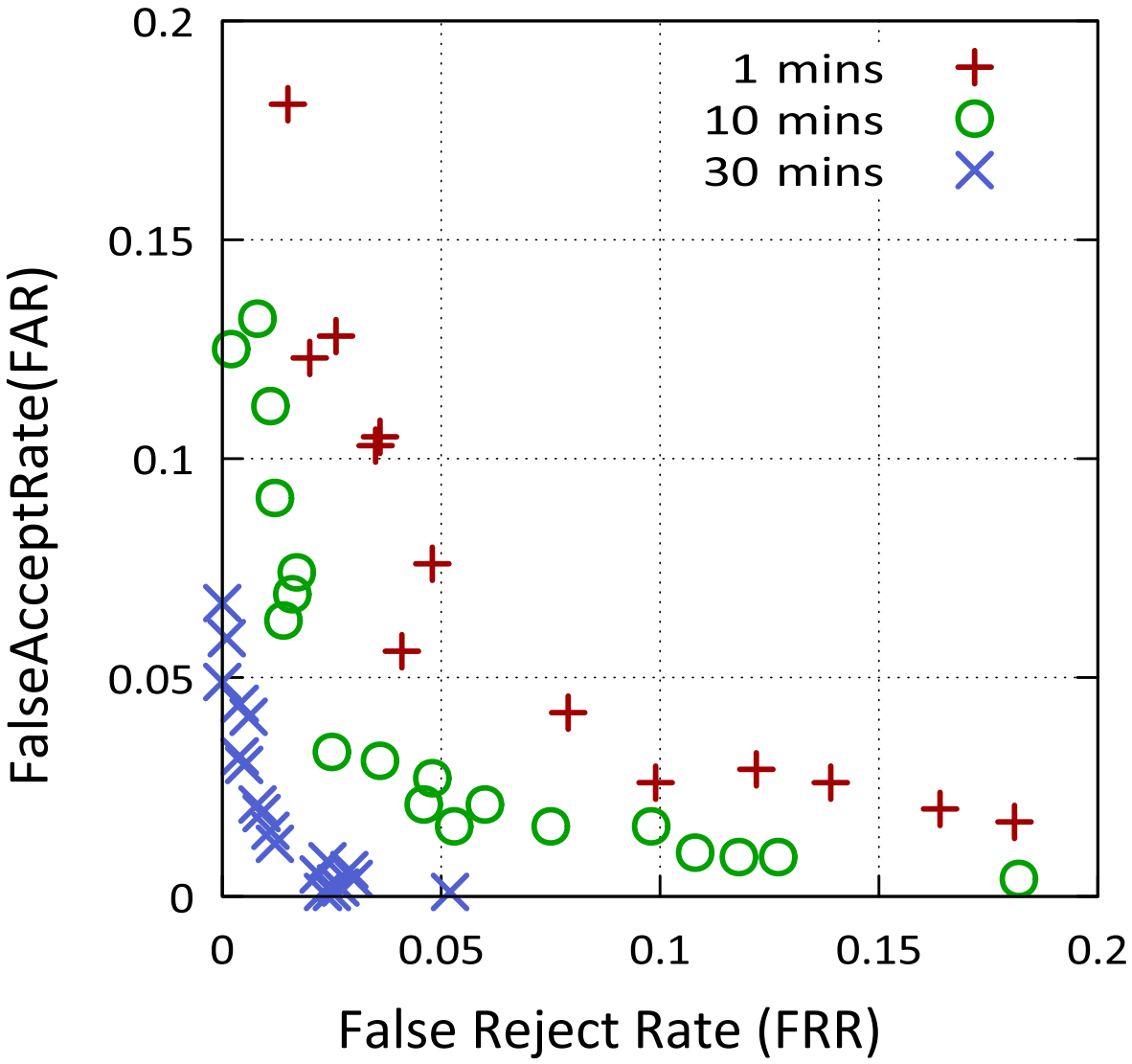

Figure 7: The performance of the fusion system with 4 classifiers on the 200 subject dataset. The ROC curve shows the tradeoff between FAR and FRR achieved by varying the threshold parameter a_0 in (4).

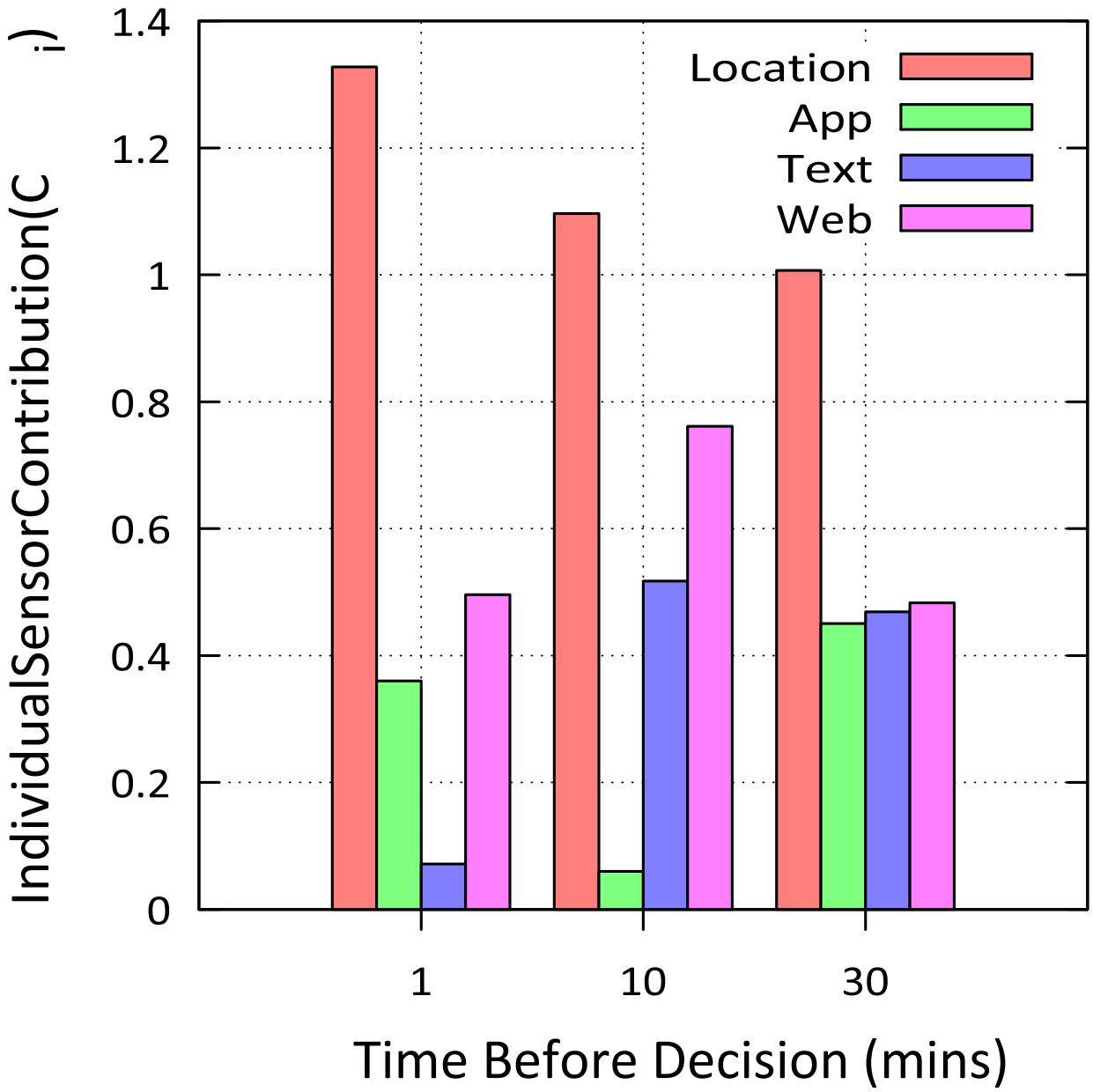

Figure 8: Relative contribution of each of the 4 classifiers computed according to (7).

6 Conclusions

In this work, we proposed a parallel binary decision-level fusion architecture for classifiers based on four biometric modalities: text, application usage, web browsing, and location. Using this fusion method we addressed the problem of active authentication and characterized its performance on a real-world dataset of 200 subjects, each using their personal Android mobile device for a period of at least 30 days. The authentication system achieved an equal error rate (ERR) of 0.05 (5%) after 1 minute of user interaction with the device, and an EER of 0.01 (1%) after 30 minutes. We showed the performance of each individual classifier and its contribution to the fused global decision. The location-based classifier, while having the lowest firing rate, contributes the most to the performance of the fusion system.

7 Disclaimer

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

[1] M. Duggan, "Cell phone activities 2013," *Cell*, 2013.

[2] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner, "Are you ready to lock?" in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014, pp. 750–761.

[3] M. Harbach, E. von Zezschwitz, A. Fichtner, A. De Luca, and M. Smith, "Itsa hard lock life: A field study of smartphone (un) locking behavior and risk perception," in *Symposium on Usable Privacy and Security (SOUPS)*, 2014.

[4] D. Van Bruggen, S. Liu, M. Kajzer, A. Striegel, C. R. Crowell, and J. D'Arcy, "Modifying smartphone user locking behavior," in *Proceedings of the Ninth Symposium on Usable Privacy and Security*. ACM, 2013, p. 10.

[5] C. Shen, Z. Cai, X. Guan, and J. Wang, "On the effectiveness and applicability of mouse dynamics biometric for static authentication: A benchmark study," in *Biometrics (ICB), 2012 5th IAPR International Conference on*. IEEE, 2012, pp. 378–383.

[6] A. Fridman, A. Stolerman, S. Acharya, P. Brennan, P. Juola, R. Greenstadt, and M. Kam, "Decision fusion for multimodal active authentication," *IEEE IT Professional*, vol. 15, no. 4, July 2013.

[7] M. O. Derawi, C. Nickel, P. Bours, and C. Busch, "Unobtrusive user-authentication on mobile phones using biometric gait recognition," in *Intelligent Information Hiding and*

Multimedia Signal Processing (IIH-MSP), 2010 Sixth International Conference on. IEEE, 2010, pp. 306–311.

[8] F. Li, N. Clarke, M. Papadaki, and P. Dowland, "Active authentication for mobile devices utilising behaviour profiling," *International Journal of Information Security*, vol. 13, no. 3, pp. 229–244, 2014.

[9] T. Sim, S. Zhang, R. Janakiraman, and S. Kumar, "Continuous verification using multimodal biometrics," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 4, pp. 687–700, 2007.

[10] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 226–239, 1998.

[11] C.-H. Chen and C.-Y. Chen, "Optimal fusion of multimodal biometric authentication using wavelet probabilistic neural network," in *Consumer Electronics (ISCE), 2013 IEEE 17th International Symposium on*. IEEE, 2013, pp. 55–56.

[12] Z. Chair and P. Varshney, "Optimal data fusion in multiple sensor detection systems," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. AES-22, no. 1, pp. 98 –101, jan. 1986.

[13] H. Saevanee, N. Clarke, S. Furnell, and V. Biscione, "Text-based active authentication for mobile devices," in *ICT Systems Security and Privacy Protection*. Springer, 2014, pp. 99–112.

[14] M. L. Brocardo, I. Traore, S. Saad, and I. Woungang, "Authorship verification for short messages using stylometry," in *Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on*. IEEE, 2013, pp. 1–6.

[15] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 2, p. 7, 2008.

[16] A. W. E. McDonald, S. Afroz, A. Caliskan, A. Stolerman, and R. Greenstadt, "Use fewer instances of the letter "i": Toward writing style anonymization." in *Lecture Notes in Computer Science*, vol. 7384. Springer, 2012, pp. 299–318.

[17] A. Fridman, A. Stolerman, S. Acharya, P. Brennan, P. Juola, R. Greenstadt, and M. Kam, "Multi-modal decision fusion for continuous authentication," *Computers and Electrical Engineering*, p. Accepted, 2014.

[18] R. Yampolskiy, "Behavioral modeling: an overview," *American Journal of Applied Sciences*, vol. 5, no. 5, pp. 496–503, 2008.

[19] M. Abramson and D. W. Aha, "User authentication from web browsing behavior." in *FLAIRS Conference*, 2013.

[20] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proceedings of the National Academy of Sciences*, vol. 106, no. 36, pp. 15274–15278, 2009.

[21] B. Sun, F. Yu, K. Wu, and V. Leung, "Mobility-based anomaly detection in cellular mobile networks," in *Proceedings of the 3rd ACM workshop on Wireless security*. ACM, 2004, pp. 61–69.

[22] J. Hall, M. Barbeau, and E. Kranakis, "Anomaly-based intrusion detection using mobility profiles of public transportation users," in *Wireless And Mobile Computing, Networking And Communications, 2005.(WiMob'2005), IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 17–24.

[23] S. Abe, *Support vector machines for pattern classification*. Springer, 2010.

[24] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 625–632.

[25] R. R. Tenney and J. Nils R. Sandell, "Decision with distributed sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-17, pp. 501–510, 1981.

[26] M. Kam, W. Chang, and Q. Zhu, "Hardware complexity of binary distributed detection systems with isolated local bayesian detectors," *IEEE Transactions on Systems Man and Cybernetics*, vol. 21, pp. 565–571, 1991.

[27] A. R. Reibman and L. Nolte, "Optimal detection and performance of distributed sensor systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-23, pp. 24–30, 1987.

List of Acronyms

DFC Decision Fusion Center

EER Equal Error Rate

FAR False Accept Rate

FRR False Reject Rate